REPORT DOCUMENTATION PAGE

AFRL-SR-AR-TR-02-

0423

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	November 19,2002	Final (01 Apr 1999 - 31 Sept 2002)

4. TITLE AND SUBTITLE
Nonmonotonic Extrapolation of Causal Relations for Knowledge-based Decision Support Using a Bayesian Network Approach

5. FUNDING NUMBERS
F49620-99-1-0211

6. AUTHOR(S)
Dr. Qiuming Zhu

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Nebraska at Omaha
Computer Science, EAB 202
6001 Dodge Street
Omahaa, NE 68182-0500

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AFOSR/NM
801 N. Randolph St., Room 732
Arlington, VA 22203-1977

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT

20030106 113

13. ABSTRACT *(Maximum 200 Words)*

This research focuses on an investigation of a new computational model for providing reliable decision support in military operations. The PI used a Bayesian network representation and a human-centered reasoning technique to extrapolate the causal relations between the available data sets stored in heterogeneous databases and the phenomena implications coherent to real world situations. The PI had developed a software agent interface for integrating the data mining and decision support operations. A nonmonotonic reasoning paradigm for derivation of causal relations was implemented in an integration of relevant software modules. The agent interface was featured with an interactive graphics setting for display and manipulation of the Bayesian Network representations, for multiple database accesses, and for belief propagations. An interactive and iterative knowledge acquisition and Bayesian reasoning system prototype was developed on top of the agent interface.

14. SUBJECT TERMS			15. NUMBER OF PAGES
			16. PRICE CODE
17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT

Table of Content

Executive Summary

This is the final report on the research project sponsored by the Air Force Office of Scientific Research (AFOSR) in its Grant NO: F49620-99-1-0211, funding from April 1st, 1999 to March 31, 2002. A 6-month no-cost extension of the project to September 30, 2002, was granted in January 2002.

The research activities have been conducted very well in achieving its original objectives. A software agent interface was developed for integrating the data mining and decision support operations. The agent interface is featured with an interactive graphics setting for display and manipulation of the Bayesian Network representation, the multiple database accesses, and the propagation of nonmonotonic reasoning. An interactive and iterative knowledge acquisition and Bayesian reasoning system is developed on top of the agent interface for providing tactical decision support in military and commercial applications. In addition, as the research work proceeds, some related issues, such as the handling of incomplete data, the handling of data with both numeric and categorical attributes, and the scale up of the data processing, etc., have emerged and addressed. Research results on these issues have expanded the original objectives.

Two faculty members, one postdoctoral researcher, and eight graduate students were supported by the grant of this research. A list of people involved in is included as appendix A.1 of this report.

The project has resulted seventeen journal and conference publications co-authored by the PI, the research assistants and graduate students. Subjects covered by the publications include "A Pseudo Nearest-Neighbor Substitution Approach for Missing Data Recovery on Gaussian Random Data Sets," *Pattern Recognition Letters*, Vol. 23, No. 13, pp. 1613-1622; "A Bayesian Network-Based Interactive and Iterative Reasoning for Decision Support System Under Uncertainty," Proceedings of the *International Conference on Artificial Intelligence*, pp. 464-470, 2002. "Finding causal patterns from frequent item sets," Proceedings of the 6th *Joint Conference on Information Systems*, pp. 442-445, 2002. "An Integrated Interactive Environment for Knowledge Discovery from Heterogeneous Data Resources," *Journal of Information & Software Technology*, Vol. 43, pp. 487-496, 2001. "Bayesian Reasoning in an Annotated Probability Space for Decision Support with Incomplete Data Set," *Proceedings of the 8th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pp. 827-834, 2000. A complete list of the publications are included as appendix A.2 of this report.

The project has also produces eight graduate dissertations at the master's level. Subjects covered by the publications include "Feature Subset Selections in Data Clustering," "User-Guided Knowledge Discovery using Bayesian Network," "Associate rule mining from incomplete and missing data records," "An Informed Interactive Query Approach for Knowledge Discovery from Heterogeneous Databases," "A Modified K-Means Algorithm For Categorical Data Clustering," etc. The theses are accessible from the University of Nebraska at Omaha Library.

1. Introduction

This is the final report on the research project sponsored by the Air Force Office of Scientific Research (AFOSR) in its Grant NO: F49620-99-1-0211, funding from April 1st, 1999 to March 31st, 2002. A 6-month no-cost extension of the project to September 30th, 2002, was granted in January 2002.

The research project is titled "Nonmonotonic Extrapolation of Causal Relations for Knowledge-based Decision Support Using a Bayesian Network Approach." It focuses on an investigation of new computational models to provide reliable decision support in military and business settings. In this project, the PI used a Bayesian network representation and a human-centered reasoning approach to the extrapolation of causal relations between the data sets stored in heterogeneous databases and their phenomena implications that are coherent to decision making in real world situations.

The original objectives of the project are:
(1) To develop a computation model for more reliable decision support in military and business settings by mining relevant knowledge patterns from multiple databases using a Bayesian network-based representation and reasoning method.
(2) To develop a software prototype for testing and validating the computational model in time- and mission-critical military and civilian operations; and
(3) To incorporate the new model and technology into educational programs in order to train computer science graduates with a broad and in-depth understanding of the technological frontiers critical to military applications.

The research activities have been conducted very well in achieving these objectives. Starting from April 1st, 1999, the beginning date of the AFOSR grant, two faculty members, one postdoctoral researcher, and nine graduate students have involved in this project with different duration of time. The PI and his team have concentrated first on the development of a software agent interface for integrating the data mining and decision support systems. The agent interface is featured with an interactive graphics setting for display and manipulation of the Bayesian Network representation, the multiple database accesses, and the propagation of nonmonotonic reasoning. Then after the interface development, an interactive and iterative knowledge acquisition and Bayesian reasoning system is developed on top of the agent interface for providing tactical decision support in military and commercial applications. In addition, as the research work proceeds, some related issues, such as the handling of incomplete data, the processing of data with both numeric and categorical attributes, and the scaling up of the data sizes and heterogeneous resources, etc., have emerged and addressed in the project. Research results on these issues have expanded our original objectives. These research activities are discussed, and results are reported, in the following sections.

Two faculty members, one postdoctoral researcher, and eight graduate students were supported by the grant of this research.

2. Status of efforts

In this project, a user-guided Bayesian network reasoning model is developed for knowledge discovery from databases. Research in the three years period has been focused on the following aspects.

(1) Development of an interactive, integrated knowledge acquisition and reasoning system model for decision support under uncertainties.
(2) Development of an interactive graphics based agent interface for supporting the implementation of the system model.
(3) Integration of the software modules developed at different stages of this project.
(4) Refining and testing of the prototype system for evaluation of the algorithms.
(5) Incorporating the new computational model and the prototype in existing courses on decision support, data warehouses and database mining, and in supervision of graduate student independent studies and theses.
(6) Applying the developed approach in real-world applications.
(7) Identifying related new research aspects for future research.

In the computational model developed in this project, a user starts from specifying the reasoning goals (decisions to be made) at the top-most level of a Bayesian network, and refines queries under the guidance of an incrementally constructed network representation. As the process of network construction goes on, the goals are refined and knowledge necessary for the reasoning to be conducted among the nodes in the network are acquired. The refinement process will get to a point where each node in the Bayesian network represent nicely an attribute available directly or indirectly from the given databases. At this nether-most layer, the system starts a data mining process to get required values fulfilled for the original goals that are related to the decision choices. On one hand, this reasoning process proceeds automatically at this point. On the other hand, the user could also further specify the values for some attributes and obtain the impacts imposed by these values on the top-most goals. These processes allow users flexibility of using the system in different decision-making environments. The software prototype was developed using Microsoft's Visual C++ programming languages in the Windows operating system.

The project has resulted seventeen journal and conference publications co-authored by the PI, the research assistants and graduate students. The project has also produces eight graduate dissertations at the master's level. These results are listed in the appendix section of this report.

3. Accomplishments/New Findings

This section will briefly outline the major accomplishments and new findings of the project in terms of major subject areas in which the research are conducted.

3.1 Knowledge discovery from heterogeneous databases

Knowledge discovery from databases (KDD) is becoming a necessary information management technology in many industrial and business practices [11]. In most cases, KDD is an integral part of database system, especially for large databases, or the databases that have very complicated correlations. Usually, it is very difficult to mine knowledge from such databases by manual queries. Even if possible, it takes lots of time and effort. So it is necessary to develop a system that can mine data automatically and report the knowledge correctly, through a KDD process. Such a system must have a through understanding of the issues regarding to the knowledge to be discovered and the diversity of the data sources.

About the knowledge: (1) the process of discovery should be automatically or semi-automatically with limited human guide. However, human-guided knowledge discovery is a necessary tradeoff; (2) the goal of such system is to find out the rules that are embedded in bulk data; and (3) as a matter of fact, it is not cared about trivial knowledge, i.e., the information mined out should be meaningful (and useful).

About the data source: (1) the data source may be very large. It is common in enterprise that the sizes of databases are more than thousands of gigabyte; (2) the data source may be distributed geographically. This kind of issue is commonly mentioned because the real world Knowledge Discovery System should deal with many data sources at the same time; (3) the data sources may be heterogeneous. The data sources could be built in different stages of database development. For example, some data is store in file system, some in legacy model database system, some in relation model database system, and some in OODB system, and some stored in web data source format.

Observing the above facts, the following two themes are followed in this research and addressed in the implementation.

1. To have a deterministic understanding of certain military combating circumstances, in the absence of clairvoyance, a decision support system needs to be equipped with various utensils that include ways of acknowledging, representing and incorporating measures of uncertainty and employing nonmonotonic inferences. To achieve this goal, a graphics interface for the software agent is developed first in the project. The interface is developed using Microsoft Visual C++ 6.0 on a Pentium microprocessor-based desk-top computer under the Windows NT (later upgraded to windows 2000) operating system.

2. It is important in a decision making process to find a way of extending the database queries to search for and locate the correlative attributes from multiple databases once an interest of event is identified. That is, a software system devoted to provide decision support must be able to query multiple heterogeneous databases to obtain necessary and concentrated information relevant to the event. In the project a multiple database access interface is also developed. The interface can simultaneously open several database tables from heterogeneous databases of different resources, and performing attribute joining, materialized views, and data aggregate operations, so as to extract information useful for the decision-supporting goal.

3.2 Bayesian networks as knowledge representation and reasoning mechanism

A Bayesian network is a graphical model for probabilistic relationships among a set of variables. It has attracted attention in the last few years, especially recently, as a possible solution for reasoning under uncertainty in artificial intelligence. Heckerman [6, 7] applied Bayesian networks to Decision Trees and implementations in large software system development. Indrawan etc. [8] used Bayesian networks as a text retrieval engine. More researches on Bayesian networks are under their way and more techniques are evolving. This research project explores an approach of applying Bayesian networks to knowledge discovery.

The unique feature of Bayesian networks in probability inference renders it a favorable tool in the implementation of knowledge discovery under uncertainty. Many successful cases of Bayesian networks applied in knowledge discovery in specific area have been reported recently. Ahn and Ezawa of AT&T Lab [1] demonstrated that by linking the BN learning model with rigorous decision-making techniques such as influence diagrams, the decision support for real-time telemarketing operations was shown to provide an intelligent decision advice. Bayesian networks have also been proven useful in practical applications such as medical diagnosis and diagnosis of mechanical failures.

Despite recent success, the Bayesian networks applications in knowledge discovery are far away from reaching its potential. Some especially challenging obstacles still exist. These include the incorporation of background and associated knowledge, the coordination of human-computer interaction, the development of efficient inference algorithms, and so on [9]. To further explore the potential of implementing Bayesian networks in knowledge discovery, the research project represents a user-guided computational model with easy user-computer communication. In this model, the user has primary responsibility for discovering knowledge representation rules. Typically, the user makes up a hypothesis according to the knowledge discovery requirement. Tests are then carried out to verify or refute that hypothesis. Based on the information retrieved, the user may then come up with a new hypothesis, or may refine the existing hypothesis, and may verify it against the source data. Thus, several iterations will be involved in successive refinement of the hypothesis, as the user tests more refined versions. This approach is suitable for decision support in organizations where the user control over the

bulk of information is a primary concern. However, this approach is also challenging and requires some kind of user training.

One of the major advantages of Bayesian networks is that it can be represented explicitly using a graph. In fact, it was in part due to this feature of graphical representation that makes BN widely considered [3, 5] as a promising tool of knowledge-discovery. However, as many researches are focusing on developing algorithm for specific application, the research on the graphical implementation is largely overlooked and remains almost blank. In this research, the PI exploited the graphical capability of Bayesian networks to facilitate the process of knowledge discovery by easing the Bayesian network construction and maintenance [2]. This is especially valuable considering a lack of a general solution of Bayesian network implementation in knowledge discovery and, when applications are more or less specifically developed. In the PI's envision, a graphical representation could provide a direct, explicit, and highly intuitive understanding of the relationship among nodes (variables), and therefore greatly improve its quality. A dynamic interaction between the graph and the users (domain expert) could be a viable mechanism to control the refinement and growth of the Bayesian network.

In the following, the accomplishments of the project that applies the Bayesian network for knowledge discovery to supporting decision-making are reported.

3.3 Bayesian networks for discovering causal relationship in KDD

In this project, a Bayesian network inference approach is used for conducting the KDD. Bayesian Network (BN), or Causal network, as a graphical model encodes probabilistic relationships among variables of interest [10]. When it is used in conjunction with statistical techniques, it exhibits advantages for data modeling. Under a causal interpretation, the links of a Bayesian network represent influential relations quantified by the conditional probability distribution values associated with each node [12], as shown by Figure 3.1. This advantage makes BN an ideal representation for combining prior knowledge and data, and can be used to learn causal relationships of database attributes in a knowledge discovery process.

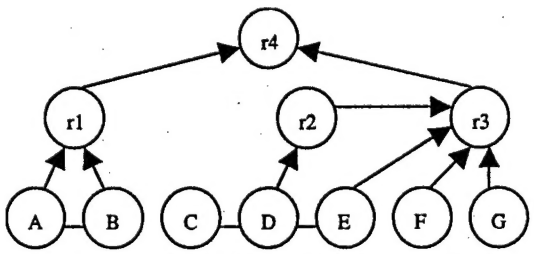

Figure 3.1 Bayesian Network Representations

Two types of nodes should be distinguished in the application of BN for data mining. These nodes are denoted as non-terminal node and terminal node. Non-terminal nodes

have at least one arrow point to it. Terminal nodes have no arrow points to it. For example, in the Figure 3.1, A, B, C, D, E, F, G are terminal nodes, and r_1, r_2, r_3, r_4 are non-terminal nodes. In this research, terminal nodes are used to represent certain attributes in databases. Non-terminal nodes are used to represent goal or sub-goals of data mining.

BN uses arrows to represent the relation between two nodes. Non-arrow lines are used to represent the causal relation that cannot be decided at the moment. That is no knowledge right now about which node directly affects the probability value of the other. Arrow line represents the causal relation between two nodes. The direction of arrow indicates the causal direction. The weight of the line indicates, in terms of probabilities, how strong the relation is.

In the real world data set, there are interrelations among different attributes. The interrelations are critical for decision support. For example, in a grocery store KDD modeling, one finds that the sale of the butter B increases as the same rate as that of bread A when the data of sales in a grocery store is analyzed. There is probably correlation existing between these two products. The grocery store could take some measures to promote the sale of them by

- Putting them together, or near each other.
- Giving coupon to one of them.
- Keeping the balance in stock between butter B and bread A. It is undesired that one product is out of stock while the other is still on sale.

On the other hand, if the sales of one product increase while the other decrease. It concludes that there is an anti-correlation between these two products. For example, the increasing sales of orange juice could result in the decreasing sales of pineapple juice. The grocery store could take some measures to deal with this situation by promoting only one kind instead of two.

Moreover, the relations are useful when some data are missing. Data missing is quite common in real world databases. For example, we cannot know the mileage in one used car because the odometer is broken. But we find the strong relation between the mileage and the age of the used car. So if we know the age of the car, we can estimate the mileage by the age of the car.

The traditional methods, for example, decision tree, cannot analyze data without knowing every detail of the data. Human must inputs the estimated value let it work. Unfortunately, the estimation made by human is sometime not so accurate that it will result in error later in its analysis. BN estimation is based on the large amount of data. It could come from one rule implicit in the data. The probability of correctness of such estimation is relatively high. In our project, the first step of mining relations is to calculate the correlation coefficient of the different columns. The resulting correlation coefficients are entered into the BN as the weight on the connections.

A KDD process with the BN representation is conducted in the following way.

(1) Extracting data and updating Bayesian networks

Automated construction of Bayesian or probabilistic networks from data is an active area of research. However, the majority of the work is not related to KDD. Our research has been focusing on the task of extracting the numbers from the database and further update these numbers. One of the core issues of our research is extracting data from the database to construct and update the Bayesian networks. To achieve this goal, we have paid attention to the following aspects.

- Approximate Inference Simulation: Approximate methods for inference have been widely used, and the most well-known one are based on simulation, i.e., using the Bayesian network to generate randomly selected instantiations of the set of parameters and then counting the number of instantiations of interest. Existing approaches have suffered various problems and a better approach is needed.

- Model Reduction: A variety of methods have been used to find or construct a simplified network as a method of attaching difficult inference problems. These methods include reduction of the domains of parameters, elimination of irrelevant arcs, local or partial evaluation, variation methods for fitting simpler parameterized models, and qualitative methods. However, none of these methods is in wide-spread use.

- Local Structure: Bayesian networks exploit value-independent conditional independence, i.e., conditional independence properties that hold at the parameter level. However, often there are conditional independence relations that hold only for specific values of a parent or child. A second kind of independence below the topological level often occurs among the parents of a parameter. Often, the effect of each parent can be modeled independently of the state of the other parents.

- Dynamic nature: Dynamic probabilistic networks are compact, factored representations of Markov processes. There has been some study of inference methods, both exact and approximate. The most common applications include projection of future values and the integration of observations over time. A stochastic approximation methods are necessary because independence relationships among variables in a static, factored state model tend to disappear when reasoning over time. As a result, exact methods are exponential in the number of variable in the state model despite the factored representation of the prior density.

(2) Building Bayesian networks for knowledge discovery

It has been noted that building a Bayesian network for a domain of application involves three major tasks:

(1) identify the variables that are of importance, along with their possible values;
(2) identify the relationships between the variables discerned and to express these in a graphical structure; and
(3) obtain the probabilities that are required for the quantitative part. The daunting question of "where do the numbers come from?" is a tough one to answer

In practice, building a probabilistic network is a process that iterates over these tasks until a network results that is deemed requisite. Despite the abundance of information, the sources seldom provide all numbers required for the quantitative part of a probabilistic network. As a consequence, the task of obtaining the numbers for a real-life application is hard and time-consuming. If a comprehensive data collection is available, the construction of both the graphical part and the quantitative part of a probabilistic network can be performed automatically. The basic idea of the former is to distill information about the relationships between the variables for the data and exploit it for constructing the network's graph. To deal with the computational complexity, there are essentially two approaches to learning the graphical structure from data: constraint-based search and Bayesian search from graphs with highest posterior probability given the data.

It is noted that many problems related to computational complexity in Bayesian network approaches are largely independent to the task of data mining. Integrating Bayesian networks with data mining has introduced new challenges to our research, but it is also a blessing, because some of the original difficulties related to Bayesian networks can be alleviated by restricting to issues related to data mining alone (rather than examining Bayesian networks in a more general perspective). The agent-based approach taken in this project has introduced a number of features to deal with the various problems, including incorporation of data mining aspects for further theoretical investigation (such as using intelligent agents for automatic data extraction from the database), as well as the development of an interactive graphics setting for display and manipulation of the Bayesian network representation.

One of the theoretical aspects we have examined is the Binary causal network model, which is a Bayesian network model also satisfying the restrictions of no more than two incoming links for a node and no more than two outgoing links for a node. Our preliminary study has shown that this model can significantly reduce the computational complexity involved in network construction and probability propagation, yet is still reasonable enough to deal with a wide range of application problems.

(3) Correlations of nodes in Bayesian network

Let X and Y be random variables with positive standard deviations σ_x and σ_y. Then the correlation between X and Y, denoted by ρ, is

$$\rho = \frac{Cov\ (X, Y)}{\sigma_x \sigma_y}$$

Cov(X,Y) is the covariance of the two variant of X and Y. For example, let X and Y has the joint probability function show in the table below.

		Value of Y		
		-1	1	P(X=x)
Value of X	-2	0.4	0.1	0.5
	-2	0.1	0.4	0.5
P(Y=y)		0.5	0.5	1.0

Note, because the means are $\mu_x = \mu_y = 0$, so

$$\rho = \frac{Cov(X,Y)}{\sigma_x \sigma_y} = \frac{E[(X-\mu_x)(Y-\mu_y)]}{\sqrt{E(X-\mu_x)^2 (Y-\mu_y)^2}} = \frac{E(XY)}{\sqrt{E(X^2)E(Y^2)}} = \frac{1.2}{\sqrt{4\times 1}} = 0.6$$

Value of ρ can range from –1 to +1 inclusive. For discrete probability functions, ρ = 1 or ρ = -1 if and only if all (x, y) pairs with nonzero probability lie along a straight line. Let us take a look at the following data.

X	3	7	1	4	6	5	10	9	8	2	6	0
Y	6	14	2	8	12	10	20	18	16	4	12	0

The correlation coefficient ρ of the above data is 1.0. This is because $Y = 2 \times X$.

Let us examine the following result. The number of sample is 10,000. We calculate $\rho_{(x,y)}$ as the following:

	1	2	3
$Y = 2X$	0.998	0.998	0.998
$Y = -2X$	0.996	0.997	0.996
$Y = X^2$	0.573	0.555	0.565
$Y = X^3$	0.752	0.748	0.751
$Y = \frac{1}{X}$	0.198	0.190	0.193
$Y = \log(X)$	0.807	0.795	0.804
$Y = e^X$	0.444	0.443	0.444
$Y = \cos(X)$	0.011	0.002	0.005
$Y = \sqrt{X}$	0.945	0.945	0.946

We can see from the above result that the correlation is a measure of linear relation.

When we calculate the correlation coefficient between two variant, the causal relation is revealed in a restricted manner. When the coefficient is large enough, that is the absolute value is near 1.0, we can say that the two variant is linear correlation. But when the coefficient is very small, that is the absolute value is near 0, we cannot say that the two

variant is not co-relational. We can use other way to find out they are co-relational or not.

In the BN representation of knowledge, the correlation coefficients between every two terminal nodes are calculated. If the abstract value of correlation coefficient is larger than one given value, we say that these two nodes (columns) have relation. In order to find out all possible relation between two different nodes, we calculate all pairs of nodes. Suppose there are *n* nodes, named as x_1 , x_2 ,....., x_i ,....., x_n. ρ_{ij} is the correlation coefficient between the node x_i and x_j. We need to calculate $C_n^2 = \frac{n!}{2(n-2)!} = \frac{n(n-1)}{2}$ different correlation coefficients.

(4) Goal driven knowledge discovery

To discover correlative attributes people tried to incorporate certain rule-refining methods in database queries in a closely user-controlled manner. A simple and useful technique is to start the knowledge-discovery process by specifying what kind of knowledge is wanted and what kind of relation is sought. These specifications may be use to indicate the directions for search and guide the analysis of the retrieved data. It may identify certain factors that are more significant or dominating than the others. In other words, users are allowed to provide criteria (or some kind of task-oriented bias) as an original goal to guide the direction of discovery. Approaches to KDD with this consideration are referred to as *goal-driven* and can be carried out through interactive queries.

A *goal* $g_i \in$ GL is defined as a proposition in an attribute extrapolation process:

$$g_i = q_i(x, p),$$

where x is a set of object and p is a set of qualifiers that applies to x. In a typical KDD process, a set of such goals are to be explored and evaluated, such as

$$G_k = \{g_{k1}, g_{k2}, g_{k3}, \ldots\}, \; G_k \subseteq G.$$

The attributes their correlation is to be extrapolated are contained or to be derived from these goals. An attribute correlation r is defined between any two goals g_{ki} and g_{kj} in G_k such that

$$r_{ij} = \gamma(g_{ki}, g_{kj});$$

where, $\gamma(g_{ki}, g_{kj})$ represents a relation between g_{ki} and g_{kj}. In general, an attribute correlation r can be defined between two subsets of G_k, such that

$$r_{IJ} = \gamma(G_{kI}, G_{kJ}),$$

where $G_{kI} \subset G_k$, $G_{kJ} \subset G_k$, and $G_{kI} \cap G_{kJ} = \varnothing$. A quantitative function can be defined on a goal, or a set of goals, such that

$$\pi[g_i] = \pi[q_i(x, p)] \text{ or}$$
$$\pi[G_k] = \pi[q_{k1}(x, p), q_{k2}(x, p), \ldots];$$

where g_{k1}, g_{k2}, ... are goals having certain correlation. The function $\pi[g_i]$ represents the strength, uncertainty, or belief of the correlation on the attributes. Let g_i and g_j be two goals. We say that g_j is a sub-goal of g_i iff $\exists r_{ij} \in R$. Let g_i be a goal and G_J be a set of goals, $g_i \notin G_J$. G_J is a sub-goal set of g_i iff

$$\forall g_j \exists r_{ij} [(g_j \in G_J) \Rightarrow \gamma(g_i, g_j)].$$

An expression indicates a possible attribute correlation with the goals is denoted as

$$\pi[g_i \mid g_j] = \pi[q_i(x, p) \mid q_j(x, p)].$$

For example, in Figure 3.2, the goal "*Software Engineer*" is represented by two sub-goals. They are "Good Education" and "*Good Experience*". It means a good software engineer should have good education and good experience. We write the relation as:

$$Software\_Engineer(x, P) = Good\_Education(x_1', P_1') \cap Good\_Experience(x_2', P_2')$$

Where "$\cap$" is an "AND" relation between the sub-goals. x_1' stands for education attributes. x_2' stands experience attributes. And P_1',P_2' stands for Good quality.

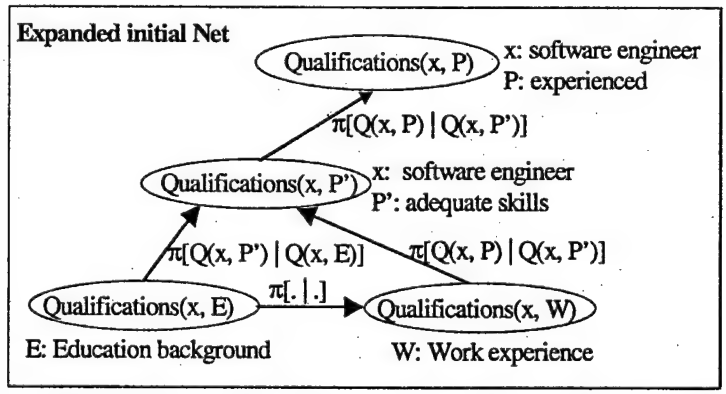

Figure 3.2 Goal-driven reasoning represented in a Bayesian network

The equation can also be written by using the $\cap$. For example, the goal G is represented by sub-goals S_1, S_2, ..., and S_n. And if all of them are "AND" relation, we represent it as G by $G = \bigcap_{i=1}^{n} W_i S_i$. The sub-goals are not always equally important to the goal, so we add weight to every sub-goal to explicitly express it. W_i is a weight for a sub-goal S_i.

In Chen and Zhu [2], it has proven that a goal driven approach provides a way for effective use of domain knowledge in data mining. The goals help identify the relevant fields of the database in an extrapolation process and excludes the irrelevant attributes. One other advantage is that, in the interactive query process, users will be allowed to

provide contingent suggestions, periodical direction corrections and intermediate result assessments, in addition to initiatives in conducting the data mining.

3.4 Interactive agent environment for knowledge discovery

In order to extract knowledge from different data sources, we have explored an agent-based approach for knowledge discovery. In this research, we concentrate on the development of a computer software agent for discovering causal relations among data objects stored in various heterogeneous databases and other resources. The goal of the research is to build an integrated agent environment that combines the data warehousing, OLAP, and KDD functionalities for supporting the knowledge discovery tasks. The system includes the tools for gathering and organizing information from multiple resources. The knowledge discovery process is based on the use of a cross-reference technique to perform informed search and a Bayesian network approach to represent the extracted rule or knowledge patterns. By incorporating aggregate data under the data warehouse environment, the system integrates the OLAP and KDD techniques to facilitate the data analysis at multidimensional data space and at different levels of abstractions.

It is generally desirable to have the process of knowledge discovery be done automatically or semi-automatically with limited human guidance. However, experiences show that a pure automatic system for knowledge discovery from multiple, large, complicated data resources is a very difficult task. Often, human-guided knowledge discovery is a necessary tradeoff. In many intelligence systems, human interaction often plays an important role to improve the performance. It has proven that human-guidance and agent-automation combined approach is a good way in KDD system because it avoids unnecessary computation so that it improves the efficiency of computation. Therefore the system we are developing is featured with an interactive interface that makes full use of human knowledge and experience to supervise the KDD smoothly and effectively.

In order to examine some important aspects of intelligent agents, a separate graduate project has been conducted for reactive agents. Experience gained from that study has shed significant light on the role of intelligent agents in database access, data extraction, and human user involvement, as well as the user-guided mining process as a whole.

(1) Algorithms for categorical data clustering and classification with bitmapping
We have continued our work of data clustering for construction of Bayesian networks in preprocessing or post-processing step for network construction. Partitioning a set of objects in databases into homogeneous groups or clusters is a fundamental operation in data mining and knowledge discovery. Clustering is a popular approach to implementing the partitioning operation. The most distinct characteristic of KDD for decision support is that it deals with very large and complex data sets (gigabytes or even terabytes). The data sets often contain millions of objects described by tens, hundreds or even thousands of various types of attributes or variables (interval, ratio, binary, ordinal, nominal, etc.). This

requires the data mining operations and algorithms to be scalable and capable of dealing with different types of attributes. In terms of clustering, we are interested in algorithms that can efficiently cluster large data sets containing both numeric and categorical values because such data sets are frequently encountered in applications. In particular, we have explored the incorporation of using bitmaps in data clustering and classification, and a new approach, called IARMBM (Influential Association Rule Mining with BitMap), has been developed, and experiments have been conducted to compare this new approach with an approach we developed earlier.

(2) Dealing with missing and imbalanced data

To allow for automated construction of a meaningful probabilistic network, the data must have been collected very carefully. Selection biases that are introduced in the data as a result of the data collection strategies used will usually have an effect on the resulting network. Unfortunately, selection biases are not easily detected in a network once it has been constructed. A common problem typically in real-life data, especially when it has been retrospectively collected, is the occurrence of missing values.. An important aspect of this task is dealing with missing data. In general, missing data recovery and handling issue is very important at data preparation step of data mining. In our last report, we identified an ongoing task would be starting the development of algorithms for handling large data sets with missing data items and attribute values, for handling the highly intertwining and categorically mixed data sets, and extracting causal relations from these data sets. Wrong approaches of missing data handling may cause the data mining system to provide false information to the users. In order to deal with missing data, we have conducted a subproject and have studied four missing data handling methods and their influences on accuracy and efficiency to Gaussian and non-Gaussian distributed data sets contain missing data values, and compared the results of Gaussain distributed data sets and non-Gaussain distributed data. The experiment used the classic k-means clustering algorithm as a tool for measurement and evaluation (Huang 2001).

Another problem with data is imbalanced data in data classification. In order to deal with this issue, we have conducted a research on classification under imbalanced data using support vector machine. The idea behind support vector machine (SVM), a very specific kind of learning machine, is that most part of data points are ignored, only a small amount of data points, called support vectors, are involved to determine the hyperplane, which separates the data points into different classes.

3.5 Mining Multiple Databases by Cross-References

Recently, most industrial corporations have built huge data warehouse that keep records of their product. It becomes more and more critical for these companies to convert this information to profit-making assets, that is, to derive commercial values from this data deposits [11]. However, the large amount of information is usually stored in separate data resources. For example, when one person applies for credit card, the credit card issuer often needs to estimate the risk of being default according to several criteria. These criteria may include:

- Whether the person ever has filed bankrupt.
- Whether the person ever has a bad credit.
- How long the person will likely stay in a well-paid job.
- How is the person's spending pattern, etc.

In order to assess the situations of above, the credit card company should access data stored in different databases. The discovery of correlative attributes from scattered data sets is a central problem in multiple database mining. It is absurd to assume that all data is stored in a single table ready for mining. In our research, we found that cross-reference and interactive query are effective methods to extract knowledge from multiple and heterogeneous databases.

We use an example to show the process involved in the cross-reference and interactive query for data mining. Let Software_Engineer(x,P) be a goal for a KDD process, where x="Software Engineer", and P="Good". The proposition Software_Engineer(x,P) means to find a knowledge pattern that gives a qualitative or quantitative definition of a "Software Engineer" as "good or desired". There are two data sources. One stores the information about the education of the software engineers; the other stores the information about the current employment status of those software engineers that is current salary, title, and so on.

- The education resource contains: *Student (SSN*, University ID, Degree, GPA, Scholarship)*
- The employment resource contains: *Software_Engineer(SSN*, Current Company ID, title, salary, times_of_promotion)*

The sign "*" means the primary key of the table.

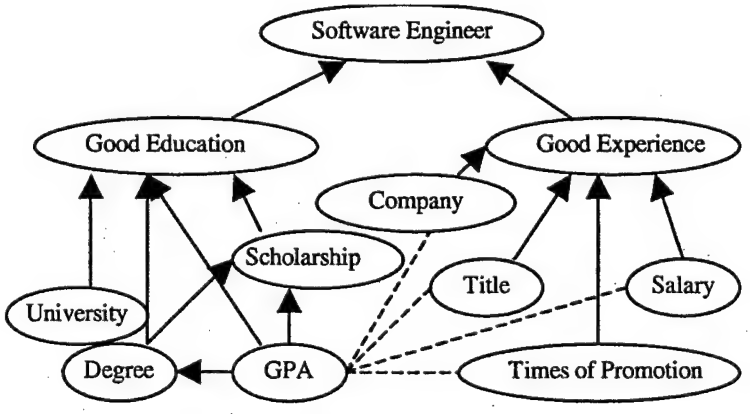

Figure 3.3 Software Engineer Problem in Bayesian network

The data source Student and data source Software-Engineer could be stored in different sites and in heterogeneous databases. If we want to know the relationship between two

attributes stored in different data sources, such as the relation between GPA and Salary (or any two attributes linked by dash line), the two different data sources must be dealt with. When relation such as GPA and Salary is mined. The system could give out all possible implicit relations, such as the relation between the GPA and Title.

One traditional and simplest way is to join all data source possibly used to a huge table. Then mine relationship in this huge table. This method is fine in many small data set applications, that is the number of the data source is not large. But in real world data mining application, the KDD system always deal with the huge among of data scattered in a great number of data sources. It is not efficient, if not impossible, to do so. Many of the attributes in the data sources do not have meaningful relations, for example University and Company.

In the interactive query, the user has the control to guide the process of mining. The user decides what relation should be mined. For example, a user want to know what software engineer's degree in his education stage can impact his later career, such as title, salary and times of promotion. If he finds that high GPA results in more times of promotion, good title and high salary in the company, and he also find (see from Figure 2.1) that there is relation between GPA and degree, so he can use cross-reference to find out that degree has relation to times of promotion, salary, and title.

3.6 Interactive Interface of IIMiner

In the developing of this project, the interactive interface was fully considered as the main feature. The interface of this system is designed to be interactive because it gives a good chance for human-guided knowledge discovery. Human interference reduces the complexity of computing greatly. In fact, the performance of system is greatly determined by the goal represented by the user. A clear and correct goal will lead to quick finding of knowledge. Interactive interface should be a preferable method to achieve.

(1) Interactive Interface

The interactive operations include the following:

1) Interactive Access of Data Sources.
   This is the first step of representing the goal. Users must see the whole detail of the data source to decide which data source will be used. In this project, it let users explore the data source freely and pick up attributes as simple as just a menu click.
2) Interactive Constructing BN Using Graphical Language.
   This is the key step of representing the goal. Users can represent the BN by nodes and arrows. Nodes describe the information of attributes and arrows describe the information among the different nodes. Users job is as easy as drawing a diagram.
3) Interactive Analysis of Retrieved Data.
   The analysis should be the core of this system. This process is designed interactively because the system is human-guided system. We hope the user can interfere with the process of the analysis. According to the definition of knowledge discovery, the

knowledge mined by the KDD system should be meaningful. As we all know, it is possible to put more intelligence into machine so that it can automatically decide which kinds of information are meaningful. But this process is extremely difficult for system because this problem is a NP problem.

This interactive feature can help users to decide which path the process should follow. Instead of calculating all C_n^2 different correlation coefficients, the system will only calculate those needed under the user's guidance.

We developed an informed interactive data mining system, called IIMiner to carry out experiment. The system implements the basic ideas described in previous sections. In addition, it also incorporates useful techniques related to heterogeneous databases and data warehouse. Figure 3.4 shows the graphical user interface of this system.

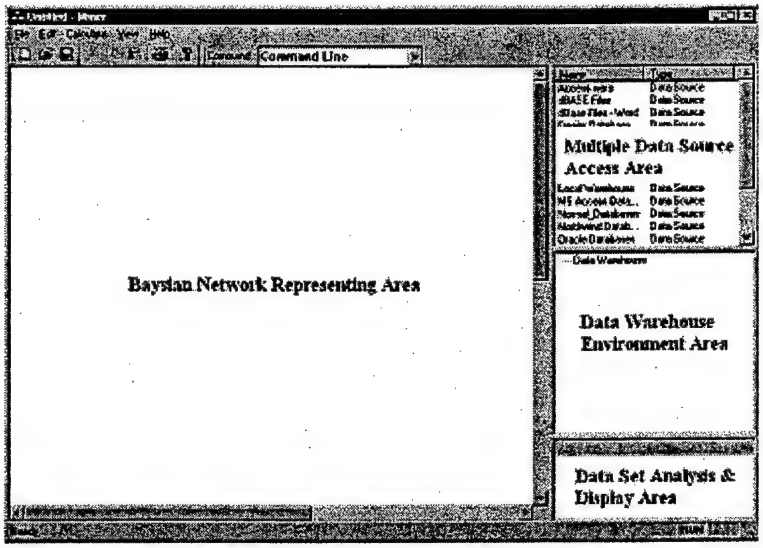

Figure 3.4 Initial State of IIMiner

There are four operating areas in IIMiner:

1. Multiple Data Source Access Area.
   This area is to explore among the heterogeneous databases. There are three levels of display.
   - Data Source Level: Display all data sources that can be access when the system is initiated. The data source could be the different sites or in different format (heterogeneous).
   - Table Level: This level is under the level of data source level. One data source contains multiple tables.
   - Attribute Level: This level is under the level of table level. One table could be composed by many attributes (columns).

2. Data Warehouse Environment Area.
   A data warehouse environment is designed to handle the heterogeneous databases. Data in this area comes from Multiple Data Source Access Area. We designed

the uniform format of the data. All different format data should be transfer to the same format of data before it can be analyzed.

3. Data Set Analysis & Display Area.
   This area is designed to show data. The data may come from specified table of a data source, or the result of mining or analysis.

4. Bayesian Network Representing Area.
   This area serves as a desktop where the user can represent his goal by Bayesian Network. We will discuss how to draw BN later.

(2) Access to heterogeneous databases

In order to access heterogeneous databases, the user should download databases that he wants to use into Data Warehouse Environment Area. The Figure 3.5 shows the table Normal1 of data source Normal-Databases in data warehouse environment.

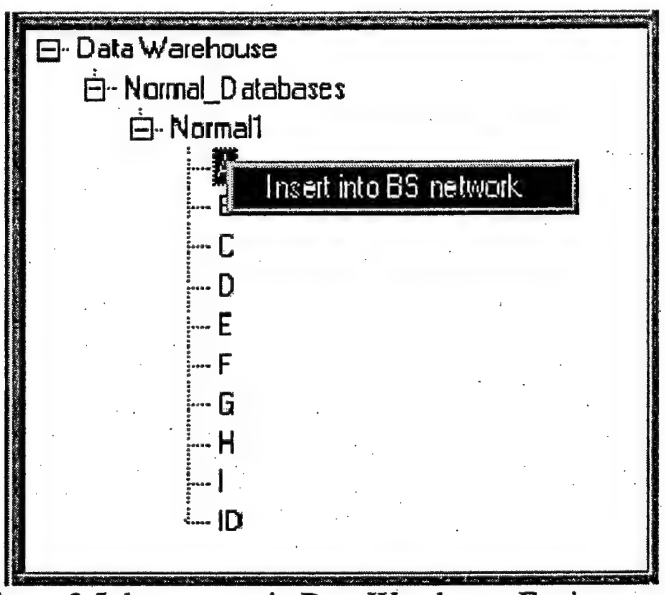

Figure 3.5 data source in Data Warehouse Environment

(3) Goal-Driven interactive Knowledge Discovery System

The process of knowledge discovery starts from the construction of an initial causal network. The initial causal network is constructed to form and represent the initial set of goal and sub-goals. These goal and sub-goals are to be expanded in the database exploration process so that the leaf nodes holds quantitative information derived from the source data. This enables the knowledge discovery process preserve a content-controllable manner. Major components of the initial Bayesian Network are:

1. An ultimate goal, which is the objective of the information retrieval process and serves as the starting point of the Bayesian network construction (the root if the causal network has a tree-like structure).

2. A set of variables (will be regarded as sub-goals) with certain relations with the ultimate goal. Here the user or the domain expert plugs in his or her own knowledge and expectations.

An initial Bayesian network could be rather coarse in terms of information content it carries. Without further refinement, it is still disjointed with the variables of relevant objects and their probabilities obtained from source data. To set up a connection, a network decomposition process is invoked. Our purpose of this process is to itemize the complexity involved in the dependency representations for the goal and sub-goals. In the decomposition process, the node connections in the Bayesian network relevant to the database attributes are identified. Intermediate nodes and links are added to the Bayesian network to set up the path from the goal and sub-goals to the leaf nodes where source data detection can be carried out. Since the network decomposition is essentially coherent with and part of the knowledge discovery process, it is integrated together with the knowledge discovery procedure. Figure 3.6 below shows a schematic diagram of the goal-driven knowledge discovery and Bayesian network management model we have developed.

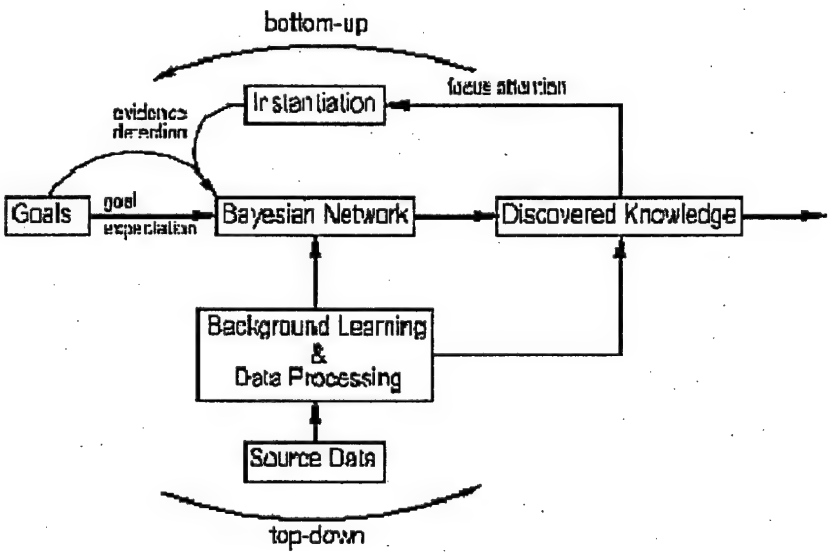

Figure 3.6. A model for user-guided knowledge discovery process

(4) Computations of Belief Propagation and Approximation

After the initialization and refinement, a tentative Bayesian network that can carry out a propagation computation consists of the following items: (a) a root node contains the ultimate goal for the knowledge discovery; (b) some leaf nodes serves as the interface to the source data; and finally (c) some intermediate or sub-goal nodes in-between the root and the leaves. There may be more than one root in the network based on the need of different application. But one of the roots, as we pointed out, must contain the ultimate goal. In a conceptual Bayesian network structure, the intermediate nodes function as a bridge where the connection between the customer desire and the source data is established.

A conceptual model has been developed in this research to represent the user-guided knowledge discovery process. This model employs an incrementally constructed Bayesian network with user-directed information. Two streams are involved in the model:

1. Top-down stream: the user enters the expectation value for the goal in a tentative Bayesian network, propagation will then be carried out to calculate the expected values for all the leave nodes.
2. Bottom-up stream: the user instantiate certain variables in the network, propagation will then be carried out to see how much the instantiation will affect the goal and the other nodes.

Both streams can continuously be managed and combined. A simple description of the main loops of the knowledge discovery process is outlined in Figure 3.7.

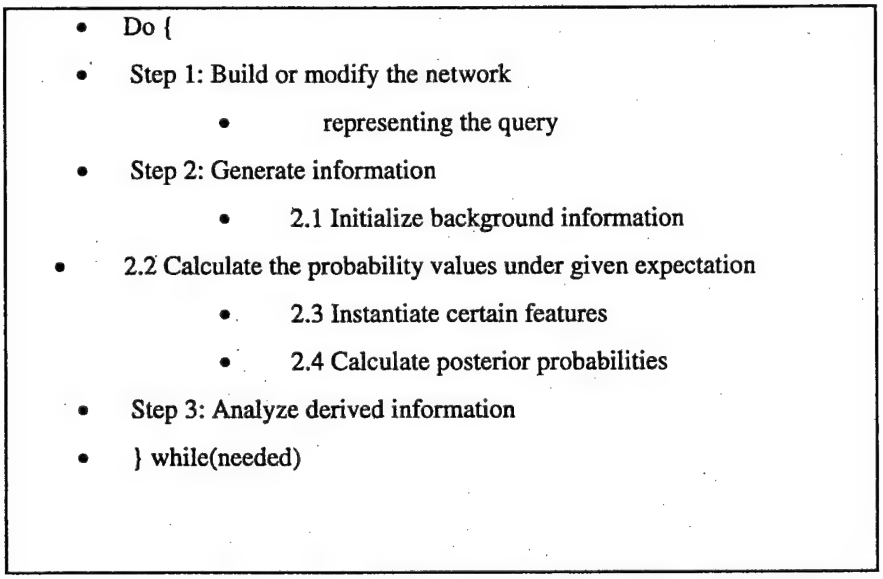

Figure 3.7. An outline of knowledge discovery process using the model

In the processes shown in Figure 3.7, step 1 is a traditional network construction and modification step. Network nodes and edges are added or deleted in Step 1. Before go over to step 2, the network should have a structure as shown in Figure 3.3. Step 2 contains the main propagation computation. The user enters background information, like conditional probabilities, in step 2.1. Following it, Step 2.2 implements the top-down calculation. The user enters an expectation value for the goal, the computation will reveal the probability value for all other nodes in the network. Upon evaluating the probability values, the user may reject certain variables or may modify the network connection according to his understanding. Furthermore, the user may focus his/her attention on certain aspects of the information by instantiate some of the variables in the network. That starts the bottom-up computation contained in step 2.3. Under the given evidence on some of the variables, the goal node will receive a new probability value after the bottom-up computation. In step 3, the user looks further into the relevance represented by the network by comparing the new value to the old expectation. More

modifications to the network may be made to better serve the knowledge discovery process and better match the domain problem as well. And the whole process repeats until no further network refinement is needed or meaningful information has already been retrieved.

In the above process, we observe two major features of our technique: interaction and iteration. We established an easy communication between the computer and the user and they work together to support the decision make. The process is a loopy refinement and testing process.

3.7 Development of the iterative software prototype

A software prototype has been developed for the interactive goal-driven knowledge discovery process. This section gives an explanation of the main frame of the software prototype.

(1) Main frame of the software prototype

Fig. 3.7 shows the main frame of the software prototype.

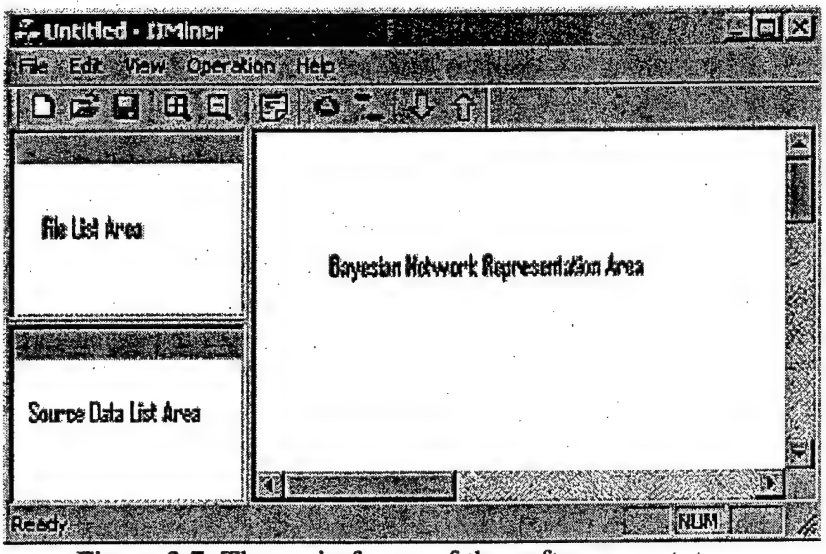

Figure 3.7. The main frame of the software prototype

Three areas have been included in the above main frame:
1) File list area: lists all available files related to the knowledge discovery
2) Source data list area: lists source data contained in a specified file
3) Bayesian Network Representation Area: represent a Bayesian network

The toolbar buttons are listed as in the figure 3.8 of following:

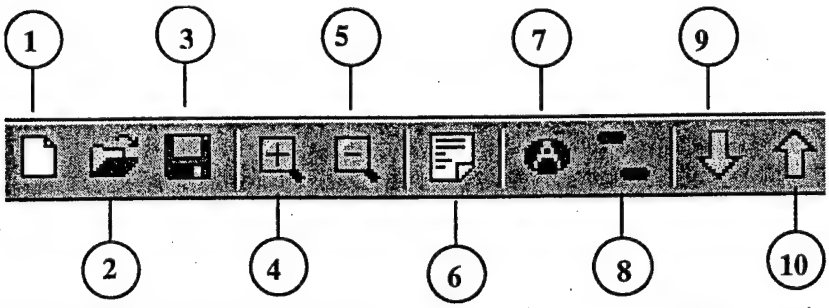

Figure 3.8. Tool Bar Buttons of the interactive Bayesian network reasoning system

The menu operators are:

1. New File: Start a new file.
2. Open File: Open an existing file.
3. Save File: Save the editing file.
4. Zoom in: Enlarge the window display area
5. Zoom Out: Shrink the window display area
6. Create Data: Create artificial data for a demonstration
7. Create Node: Add a node to the current Bayesian network.
8. Create Link: Create a new link.
9. Top-down computation: Calculate the probability of other nodes under the given probability of the root nodes.
10. Bottom-up computation: Calculate the probability under given evidence to certain nodes.

When clicked, the *Insert Node* dialog shown in Figure 3.9 will be triggered. *Open Library* button opens a related knowledge base if available. All nodes in the library will be listed in *Library Nodes* list box. To reveal the relationship among the library nodes, type a node name in the *Related to* text box and click *Show Related* button. In Fig. 5, all nodes related to Node *Department* in the given knowledge base are listed in the *Related Nodes* list box. To add a node, one can either type in a new node name in the *Node Name* text box or select one node in the *Library Nodes* or *Related Nodes* list. Figure 3.9 selects a *Working Experience* node form the *Library Nodes* list.

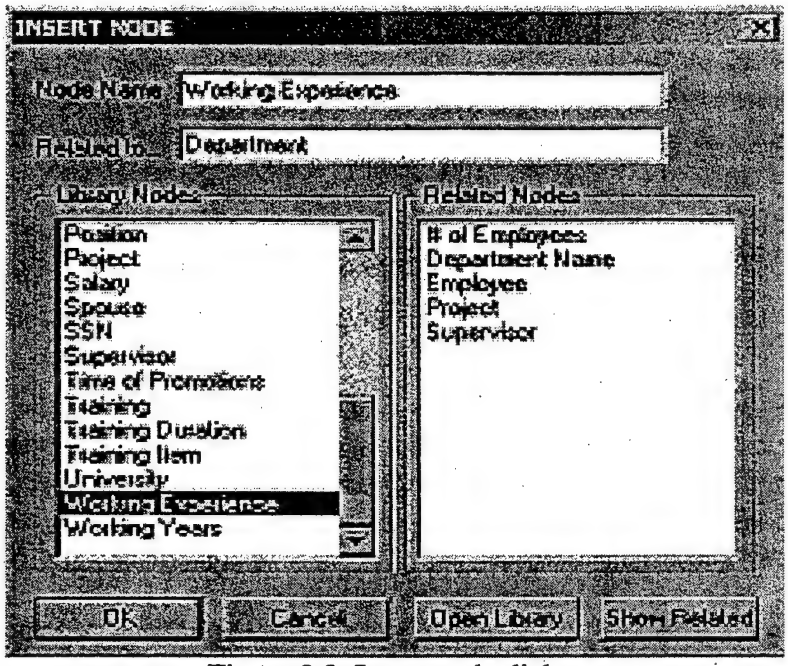

Figure 3.9. Insert node dialog

When an operation of "Create a new link" is picked for the two selected nodes, three causal relations are provided as shown in Figure 3.10, for users to select the direction of the link to establish on the nodes.

Figure 3.10. Selection buttons for link directions

(2) Interactive activity for data input and information check
To input or check the conditional probability value, double click on the link to trigger a link state dialog. To input or the probability value for a node, double click on the node opens a node state dialog.

(3) Software Engineering Demonstration
In this section, we consider a decision support system where the user is interested in discovery of some common features of qualified candidates for a software engineering position. The source data given to the user consists of employment information for previous qualified software engineers including their educational background, working history and other related information. One can carry out the goal-driven process in the following manner. Let Software Engineer be a goal for a knowledge discovery process, whose purpose is to find out the requirement under which a candidate could be a good match for the goal or if given a candidate's information, how qualified this candidate is for such a position. An initial network that consists of the goal and the sub-goals for this application is given in Figure 3.11.

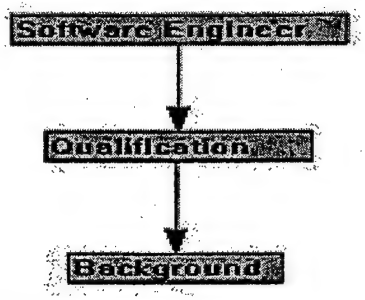

Figure 3.11. The initial network of the software engineer demonstration

We look at the source data related to the sub-goals: background. Since the source data contains information about the employer's education and working background, we then add two more node, education and working experience to the network. Keeping expanding the initial network by considering the source data contents at the meantime, we obtain the network in Figure 3.12.

As we can see from Figure 3.12, the ultimate goal for this application is contained in one of the root nodes: Software Engineer. The leaf nodes are: Certificates, training times, degree, GPA, time of promotion, working years. Qualitative information related to these nodes can be obtained by processing the source data. Other nodes in the networks serve a connection between the goal and the source data.

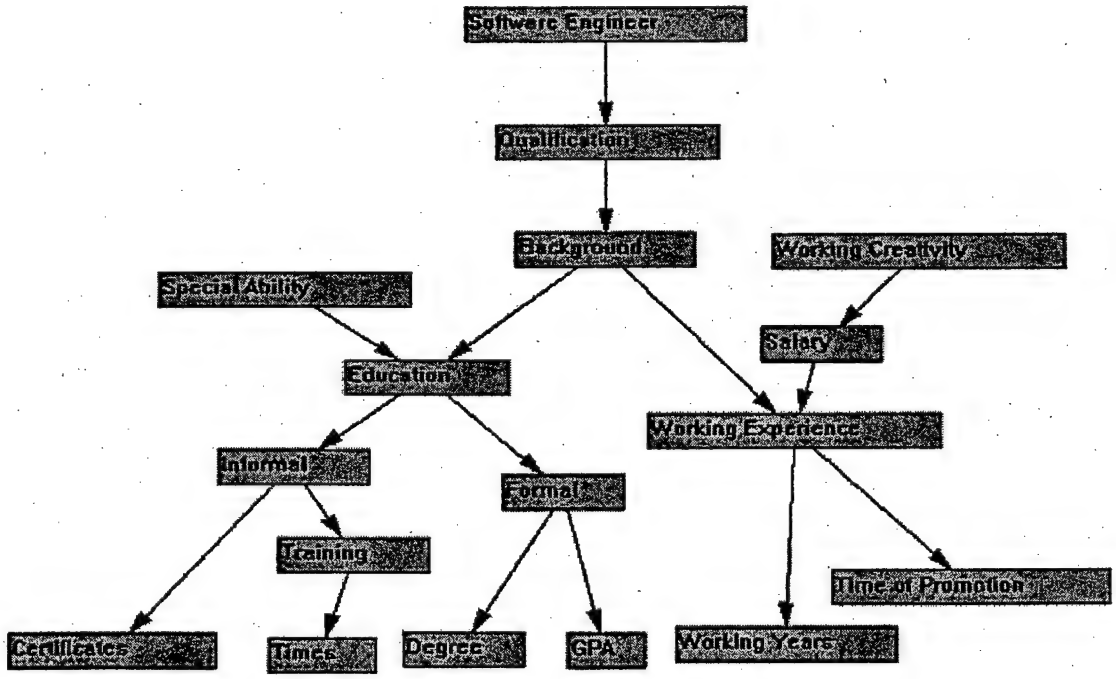

Figure 3.12. The Bayesian network structure of the software engineer demonstration

The user now enters the associated conditional probability values, input an expectation probability value for all the root nodes, and do a top-down propagation. The computation then reveals the required values for each of leave node as one shown in Figure 3.13.

When examining these values, the user may determine to remove the node Training Times since it doesn't appear to be crucial for the goal. Or the user may also want to step further to test how much the GPA will effect the qualification. Then he instantiate the GPA node, start a bottom up propagation. The probability value for the goal changed from 0.8 to 0.61, which probability tells that the GPA did play an essential role.

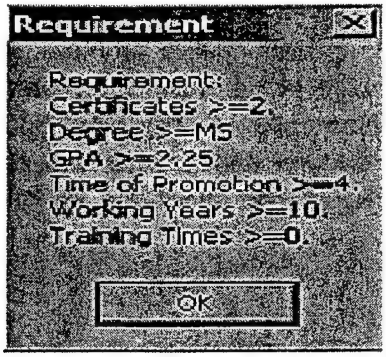

Figure 3.13. The required values for each leaf nodes after the top-down computation

4. Summary and Conclusions

The problem of mining multiple databases for discovering meaningful knowledge patterns in a specific problem domain has been addressed by the general tasks of KDD research. However, the advance of retrieving the coherent attribute relations has been hampered many years by the problems of lacking a viable and integrated approach. For example, it is easy to find many (more than needed) statistically reasonable results that are possibly insignificant or redundant in a conventional data mining process. It is necessary to have some means that identify the needed attributes of those personal qualifications, to establish a coherent relationship among the attributes, and then to have a clear picture (a knowledge pattern) of what kind of relations are looking for. The establishment of the knowledge pattern requires to extracting the correlated attributes probably from several databases involved. The work described in this project provides one solution for these types of problems. The accesses of multiple and heterogeneous data resources is considered as a main issue in this project. From our research, we find that cross-reference and interactive query are effective approaches to extract knowledge from multiple databases. Once an attribute of interest is identified, it is effective to locate other relevant attributes and extend the database query to retrieve the necessary information. This in consequence allows the system to make interactive queries according to the results from cross-references among the entries received, thus speedy up the entire KDD process.

In this project, we have developed a methodology based on a model for an interactive process of knowledge discovery. We have also developed a software prototype to support this methodology. The IIMiner system provides a convenient ways for users to interact with the KDD processes and to guide the search of information in the formation

of useful knowledge patterns. It has proven that human-guidance and agent-automation combined approach is a good way in constructing KDD systems because it avoids unnecessary computational burdens and improves the efficiency of processes. We demonstrated the use of the software prototype by exploring a software engineer application example. Since this knowledge discovery approach is interactive, the user can always provide his or her intention to guide the search in a timely fashion. Consequently, the scope of directions to consider can be reduced, unwanted knowledge patterns can be excluded, and the overabundance problem can be remedied. When the discovery process iterates, unwanted or unrelated information are filtered out and the knowledge pattern shrank to be more realistic. This new approach is most suitable to the cases of decision support where the user knows what he wants and is able to refine his goal step by step, that is, a kind of control of the knowledge discovery process is desirable to the users.

References

[1] Ahn, Jae-Hyeon; Ezawa, Kazuo J., "Decision support for real-time telemarketing operations through Bayesian network learning", *Decision Support Systems*, v 21, n 1, pp.17-27, Sep, 1997

[2] Chen, Z., and Zhu, Q., "Query Construction for User-guided Knowledge Discovery in Databases", *Journal of Information Sciences*, No. 109, pp. 49-64, 1998

[3] Cook, D. J.; Holder, L. B., "Graph-Based Data Mining," *IEEE Intelligent Systems & Their Applications*, Vol. 15, No. 2, pp.32-39, 2000

[4] Cooper, G., "Computational complexity of probabilistic inference using Bayesian belief networks," *Artificial Intelligence*, Vol.42, pp. 393-405, 1990

[5] Heckerman, D., Bayesian networks for knowledge discovery, *AI Magazine 15(3)*, pp. 273-305, 1994.

[6] Heckerman, D.; Breese, J. S.; Rommelse, K., "Decision-Theoretic Troubleshooting," *Communication of the ACM*, Vol. 38, No. 3, pp. 49-57, 1995

[7] Heckerman, D.; Wellman, M. P., "Real-World Application of Bayesian Networks," *Communication of the ACM*, Vol. 38, No. 3, pp. 25-41, 1995

[8] Indrawan, M.; Ghazfan, D.; Srinivasan, B., "Using Bayesian Network as Retrieval Engines," *Proceedings 15th Text Retrieval Conference*, pp. 437-443, 1996

[9] Neapolitan, E. R., "Probabilistic Reasoning in Expert Systems: Theory and Algorithms", *A Wiley-Interscience Publication*, 1990

[10] Pearl, J., Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, Morgan Kaufmann Publishers, Inc., 1988.

[11] Piatesky-Shapiro, G., and W.J. Frawley, W. J., (Eds.), Knowledge Discovery in Databases, *AAAI/MIT Press*, Menlo Park, CA, 1991.

[12] Zhu, Q., and Chen, Z., Knowledge Discovery from Databases with the Guidance of Causal Network, Z. Ras and A. Skowron(eds.), Foundations of Intelligent System, 10th Int'l Symp., ISMIS'97, Lecture Notes in Artificial Intelligence1352, pp. 401-410, 1997

Appendix

A.1. Personnel Supported

Faculty:

	Name	Position	Affiliation	Period
1	Zhengxin Chen	Professor	Department of computer science, University of Nebraska at Omaha	March 1999 - September 2002
2	Qiuming Zhu	Professor	Department of computer science, University of Nebraska at Omaha	March 1999 - September 2002

Postdoctoral researcher:

	Name	Position	Affiliation	Period
1	Li Xiao	Researcher	Department of computer science, University of Nebraska at Omaha	July 1999 - June 2002

Graduate Student:

	Name	Position	Affiliation	Period
1	Miao Chen	Graduate student	Department of computer science, University of Nebraska at Omaha	May 1999 - June 2000
2	Xihua Zhang	Graduate student	Department of computer science, University of Nebraska at Omaha	May 1999 - June 2000
3	Xiaolu Huang	Graduate student	Department of computer science, University of Nebraska at Omaha	January 2000 - December 2001
4	Rong Fan	Graduate student	Department of computer science, University of Nebraska at Omaha	May 2000 - June 2001
5	Xiaobu Wu	Graduate student	Department of computer science, University of Nebraska at Omaha	June 2000 - May 2001
6	Jie Deng	Graduate student	Department of computer science, University of Nebraska at Omaha	September 2000 - May 2000
7	Arun Dhulipala	Graduate student	Department of computer science, University of Nebraska at Omaha	September 2001 - August 2002
8	Hongmei Cui	Graduate student	Department of computer science, University of Nebraska at Omaha	May 2002 - August 2002

The PI has secured cost sharing support from the host institution, the University of Nebraska at Omaha (UNOmaha), in the form of 0.25 FTE assigned research time through academic salaries and benefits spending for the PI to conduct this research during academic months from August 1999 to April 2002. The PI has also successfully obtained support from the Center for Management Information Technology (CMIT) at UNOmaha for one graduate assistant in a contribution of over $19,800 (stipend plus tuition) from May 1999 to April 2000. The College of Information Science and Technology at UNOmaha has also supported one graduate assistant for the academic years of 1999 to 2002, in the total estimated amount of $19,800 (stipend plus tuition) per year.

A.2. Publications

(1) X. Huang and Q. Zhu, "A Pseudo Nearest-Neighbor Substitution Approach for Missing Data Recovery on Gaussian Random Data Sets," *Pattern Recognition Letters*, Vol. 23, No. 13, pp. 1613-1622, 2002.

(2) Y. Sun, and Q. Zhu, "An iterative initial-points refinement algorithm for categorical data clustering," *Pattern Recognition Letters*, Vol. 23, No. 7, pp. 875-884, 2002.

(3) X. Wu, Q. Zhu, "A Hierarchical Algorithm for Clustering Class-imbalanced Datasets," Proceedings of the *International Conference on Artificial Intelligence*, IC-AI 2002, pp. 457-463, Las Vegas, June 24-27, 2002.

(4) J. Deng, "A Bayesian Network-Based Interactive and Iterative Reasoning for Decision Support System Under Uncertainty," Proceedings of the *International Conference on Artificial Intelligence*, IC-AI 2002, pp. 464-470, Las Vegas, June 24-27, 2002.

(5) L. Xiao, Z. Chen, Q. Zhu, "Finding causal patterns from frequent item sets," Proceedings of the 6th *Joint Conference on Information Systems* (JCIS2002), pp. 442-445, Research Triangle Park, North Carolina, March 8-13, 2002.

(6) X. Zhang, Z. Chen, Q. Zhu, "Mining influential association rules," Proceedings of the 6th *Joint Conference on Information Systems* (JCIS2002), pp. 490-493, Research Triangle Park, North Carolina, March 8-13, 2002.

(7) M. Chen, Q. Zhu, and Z. Chen, "An Integrated Interactive Environment for Knowledge Discovery from Heterogeneous Data Resources," *Journal of Information & Software Technology*, Vol. 43, pp. 487-496, 2001.

(8) Q. Zhu, "Research on multi-subclass modeling for pattern recognition in complex spaces," Book chapter, *Recent Research Developments in Pattern Recognition*, 2001.

(9) Q. Zhu, "A multiple hyper-ellipsoidal subclass model for an evolutionary classifier," *Journal of Pattern Recognition*, 2000.

(10) Q. Zhu, "Bayesian Reasoning in an Annotated Probability Space for Decision Support with Incomplete Data Set," *IPMU2000: The 8th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, July 2000, Madrid, Spain

(11) M. Chen, Q. Zhu, and Z. Chen, "An Informed Interactive Query Approach for Knowledge Discovery from Heterogeneous Databases," *The 2000 International Conference on Artificial Intelligence (IC-AI'2000)*, pp. 15-22, June 2000, Las Vegas, USA.

(12) Y. Sun, Q. Zhu, and Z. Chen, "A Modified K-means Algorithm for Categorical Data Clustering," *The 2000 International Conference on Artificial Intelligence (IC-AI'2000)*, pp. 31-37, June 2000, Las Vegas, USA.

(13) L. Xiao, Z. Chen and Q. Zhu, "Storing and Querying XML Data for Bayesian Network Inferences," *The 2000 International Conference on Artificial Intelligence (IC-AI'2000)*, pp. 7-14, June 2000, Las Vegas, USA.

(14) L. Xiao, Z. Chen, and Q. Zhu, "A Bayesian Approach to Mining Causal Relationship Patterns from Frequent Data Sets" *PADD'2000: The Fourth International Conference and Exhibition on The Practical Application of Knowledge Discovery and Data Mining*, April 2000, UK.

(15) M. Chen, Q. Zhu, and Z. Chen, "Informed Interactive Query for Knowledge Discovery from Heterogeneous Databases," *PADD'2000: The Fourth International Conference and Exhibition on The Practical Application of Knowledge Discovery and Data Mining*, April 2000, UK.

(16) M. Khanijo, Z. Chen and Q. Zhu, "Spatial Data Access Methods Using Branch-grafted R-trees," *Proceedings of 12th International Conference on Systems Research, Informatics and Cybernetics,* Aug. 2000.

(17) Q. Zhu, and Z. Chen, "Mining Multiple Databases by Cross-referencing and Interactive Queries," *ICICS'99, the second International Conference on Information, Communications and Signal Processing*, Session 3D4, #4, Dec. 1999.

A.3. Theses

(1) Qinglu Kong, "A mini-max Ball/Ellipsoid Clustering Approach to the Modeling of Large and Complicatedly Distributed Data Sets." September 2002. Available at the University Library, University of Nebraska at Omaha.

(2) Xiaobo Wu, "Feature Subset Selections in Data Clustering," July 2002. Available at the University Library, University of Nebraska at Omaha.

(3) Jie Deng, "User-Guided Knowledge Discovery using Bayesian Network," July 2002. Available at the University Library, University of Nebraska at Omaha.

(4) Hongyang Tan, "Associate rule mining from incomplete and missing data records," March 2002. Available at the University Library, University of Nebraska at Omaha.

(5) Rong Fan, "Support Vector Machine for Pattern Classification," May 2001. Available at the University Library, University of Nebraska at Omaha.

(6) Xiaolu Huang, "An Experimental Study of Randomly Distributed Missing Data Handling Approaches for Data Mining," May 2001. Available at the University Library, University of Nebraska at Omaha.

(7) Maio Chen, "An Informed Interactive Query Approach for Knowledge Discovery from Heterogeneous Databases," June 2000. Available at the University Library, University of Nebraska at Omaha.

(8) Ying Sun, "A Modified K-Means Algorithm For Categorical Data Clustering," May 2000. Available at the University Library, University of Nebraska at Omaha.

A.4. Interactions / Transitions

Participation/presentations at meetings, conferences, seminars, etc.

- The International Conference on Artificial Intelligence, IC-AI 2002, Las Vegas, June 24-27, 2002.
- The AAAI/KDD/UAI-2002 Joint Workshop on Real-time Decision Support and Diagnosis Systems, Edmonton, Alberta, Canada, July 2002.
- International Conference on Artificial Intelligence, IC-AI 2002, Las Vegas, June 24-27, 2002.
- The 6th Joint Conference on Information Systems (JCIS2002), Durham, NC, March 8-10, 2002.

- PADD'2000: The Fourth International Conference and Exhibition on The Practical Application of Knowledge Discovery and Data Mining, April 2000, UK.
- IPMU2000: The 8th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, July 2000, Madrid, Spain.
- The 2000 International Conference on Artificial Intelligence (IC-AI'2000), June 2000, Las Vegas, USA.